Releasd Infrastructure Review

Objective

The purpose of this document is to investigate and highlight any potential areas of improvement and/or cost optimisation within the existing Releasd infrastructure.

After our initial discussion it was made clear that there are three key areas to consider:

· Reduction of Operational Expenditure – The expenses the business incurs to ensure business continuity. These are currently running at approximately £3,500 and you feel like this can be reduced.

· Update existing software/hardware associated with running services – Some of the existing infrastructure was provisioned over 12 month ago and may require security updates. There is scheduled maintenance to upgrade the MySql database engine and any updates should ideally coincide with this maintenance window.

· Any further recommendations – To look at any further areas of improvement that could be used to improve efficiency, reduce cost or increase security.

Infrastructure Overview

Introduction

It is important to consider the five pillars of the Well Architected Framework when planning or reviewing infrastructure.

These five pillars are:

· **Operational Excellence** – This includes the ability to support development and run workloads effectively, gain insight into their operation, and continuously improve supporting processes and procedures to deliver business value.

· **Security** – The business must have the ability to protect data, systems and assets.

· **Reliability** – The workload must have the ability to perform its intended function correctly and consistently when expected to do so.

· **Performance Efficiency** – Computing resources should be used efficiently to meet system requirements and maintain that efficiency as business demand and technologies change.

· **Cost Optimisation** – Systems that deliver business value must do so at the lowest price point.

Current Infrastructure

The existing infrastructure is documented in the google documents file. I have audited the AWS account and listed the existing infrastructure inside this document.

Virtual Private Cloud

A Virtual Private Cloud is essentially your own private datacentre located in the existing AWS Infrastructure.

Resource Type	Name	ARN ID	Region
VPC	released-production	vpc-e6532b81	Eu-west-1 (Ire)
VPC	released-staging	vpc-a65a23c1	Eu-west-1 (Ire)
VPC	Behaviour Tracking	vpc-0dd44e5172569bfeb	Eu-west-1 (Ire)
VPC	-	vpc-656ce50d	Eu-west-1 (Ire)
VPC	Pdf-converter? + CNAME-redirector	vpc-0a4ef16f	Eu-west-1 (Ire)
VPC	Released-node-utilities-vpc	vpc-0bf93eeadc2a439c5	Eu-west-2 (Lon)
VPC	Vapor-network-1568212218	vpc-0fe40f507dd523194	Eu-west-2 (Lon)
VPC	Releasd-pdf-vpc	vpc-0e52b263e3a50e2f9	Eu-west-2 (Lon)
VPC	Releasd-vpc	vpc-0afb6fc918ec36c5e	Eu-west-2 (Lon)
VPC	-	vpc-a858a4c1	Eu-west-2 (Lon)

Subnet

A VPC is split between a number of Subnets, a smaller network within the VPC usually for a specific purpose or a logical set of resources.

Resource Type	Name	ARN ID	Region
SubNet	–	subnet-adbf2cda	Eu-west-1 (Ire)
SubNet	–	subnet-7b6ce513	Eu-west-1 (Ire)
SubNet	releasd-production-public-eu-west-1a	subnet-1f94ea56	Eu-west-1 (Ire)
SubNet	releasd-staging-public-eu-west-1b	subnet-051cac5e	Eu-west-1 (Ire)
SubNet	releasd-staging-internal	subnet-a5cb7bfe	Eu-west-1 (Ire)
SubNet	–	subnet-7680eb13	Eu-west-1 (Ire)
SubNet	–	subnet-05e7b1140bd4843c2	Eu-west-1 (Ire)
SubNet	releasd-production-public-eu-west-1c	subnet-18a4d37f	Eu-west-1 (Ire)
SubNet	releasd-production-public-eu-west-1b	subnet-a0de6dfb	Eu-west-1 (Ire)
SubNet	–	subnet-06299c5f	Eu-west-1 (Ire)

SubNet	–	subnet-0ed60da2d37040692	Eu-west-2 (Lon)
SubNet	Public Subnet 2	subnet-0077f58a016e0b970	Eu-west-2 (Lon)
SubNet	–	subnet-1a0eee61	Eu-west-2 (Lon)
SubNet	–	subnet-28c9cc62	Eu-west-2 (Lon)
SubNet	Public Subnet 1	subnet-019b9d73fb09af739	Eu-west-2 (Lon)
SubNet	–	subnet-18b57071	Eu-west-2 (Lon)
SubNet	–	subnet-070a6ec5299993606	Eu-west-2 (Lon)
SubNet	Private Subnet 1	subnet-02a40dfadb58521d1	Eu-west-2 (Lon)

Internet Gateway

A scalable resource that allows communication between your VPC and the internet.

Estimated Cost	Name	ARN ID	Region
£24.47 / month	–	igw-067cff5b8850a6ba8	Eu-west-2 (Lon)
£24.47 / month	–	igw-0a2c167a0df6b9eae	Eu-west-2 (Lon)
£24.47 / month	–	igw-0d47b067d8e426730	Eu-west-2 (Lon)
£24.47 / month	node-utilities	igw-0fa8657c1815ec685	Eu-west-2 (Lon)
£24.47 / month	–	igw-4b50a122	Eu-west-2 (Lon)
£24.47 / month	–	igw-0c8fa1a54f6f9a557	Eu-west-1 (Ire)
£24.47 / month	releasd-production	igw-0f3d336b	Eu-west-1 (Ire)
£24.47 / month	–	igw-666ce50e	Eu-west-1 (Ire)
£24.47 / month	releasd-staging	igw-a48e81c0	Eu-west-1 (Ire)
£24.47 / month	–	igw-d6c012b3	Eu-west-1 (Ire)
£244.70 / month **£2936.40 / annum**			

Elastic Compute Cloud

These are your secure, resizable compute capacity (web servers etc) located within the VPC.

Estimated Cost	Name	Size	Region
£61.41 / month	releasd-staging-web-1-vpc	m3.medium	Eu-west-1 (Ire)
£54.35 / month	releasd-staging-worker-1-vpc	m4.large	Eu-west-1 (Ire)
£54.35 / month	releasd-production-worker-1	m4.large	Eu-west-1 (Ire)
£489.18 / month	production-web-2 20191114	m3.2xlarge	Eu-west-1 (Ire)
£3.72 / month	CNAME	t2.nano	Eu-west-1 (Ire)
£489.18 / month	production-web-1 20191114	m3.2xlarge	Eu-west-1 (Ire)
£412.43 / month	releasd-pdf-service	m5.4xlarge	Eu-west-2 (Lon)
£1,564.62 / month **£18,775.44 / annum**			

RDS

These are your secure, resizable compute capacity (web servers etc) located within the VPC.

Estimated Cost	Name	Size	CPU
£73.95 / month	encrypted-staging	db.t2.large	1.33%
-	payments-db-cluster	1 instance	-
£18.49 / month	payments-db	db.t2.small	11.69%
£435.23 / month	releasd2	db.m3.xlarge	1.92%
£217.61 / month	releasd2replica	db.m3.xlarge	0.67%
£745.28 / month **£8943.36 / annum**			

Cost Comparisons

There are ways to immediately reduce cost without reducing capacity:

1. Upgrade existing servers from m3 (previous generation) to m5 – these run on more up-to-date hardware and are in most cases more cost effective than their predecessors.
2. Utilise reserved instance pricing.

Below is an example of the estimated monthly running cost for the current EC2 instances in the current configuration and then with the updated configuration (running on newer generation hardware) and the potential cost saving for making this switch.

Current Estimated Cost	Current Size	New Size	New Estimated Costs
£61.41 / month	m3.medium	m3.medium	£61.41 / month
£54.35 / month	m4.large	M5.large	£52.16 / month
£54.35 / month	m4.large	M5.large	£52.16 / month
£489.18 / month	m3.2xlarge	M5.2xlarge	£228 / month
£3.72 / month	t2.nano	t2.nano	£3.72 / month
£489.18 / month	m3.2xlarge	M5.2xlarge	£228 / month
£412.43 / month	m5.4xlarge	m5.4xlarge	£412.43 / month
£1,564.62 / month **£18,775.44 / annum**		33% Cost Saving	**£1,037.88 / month** **£12,454 / annum**

We could apply the same logic to the RDS databases.

Estimated Cost	Current Size	New Size	CPU
£73.95 / month	db.t2.large	db.t2.large	£73.95 / month
-	1 instance	1 instance	-
£18.49 / month	db.t2.small	db.t2.small	£18.49 / month
£435.23 / month	db.m3.xlarge	db.m5.xlarge	£410.94 / month
£217.61 / month	releasd2replica	db.m5.xlarge	£205.47 / month
£745.28 / month **£8943.36 / annum**		4.8% Cost saving	**£708.85 / month** **£8506.20 / annum**

Simply implementing this change would give a cost saving of approximately 24% (£6760 per annum / £563 per month).

Another option is to scale down our resources according to historical demand. Looking at the telemetric data from CloudWatch, DataDog and New Relic I believe the following changes can be made:

· **Releasd-staging-web-1-vpc** & **Released-staging-worker-1-vpc**

If these instances are only required when developing new features to the existing application, I would suggest these be stopped until needed. This will significantly reduce the running costs of these instances.

Estimated saving: £115.76 per month / £1389.12 per annum

We could also stop the production RDS server **encrypted-staging** which would give a further reduction of approximately: £73.95 per month / £887.40 per annum.

This action alone would provide a total saving of:

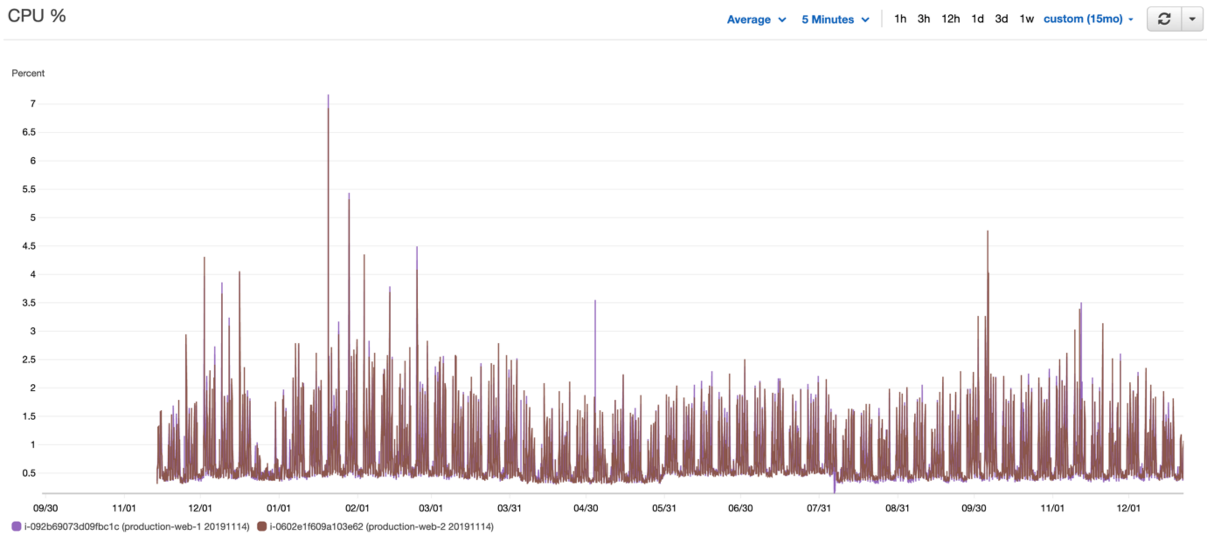
£189.71 per month
£2276.52 per annum

· **Production-web-1** & **Production-web-2**

The current specs of these servers are the following:

30GB Ram
8 vCPUs
160GB storage

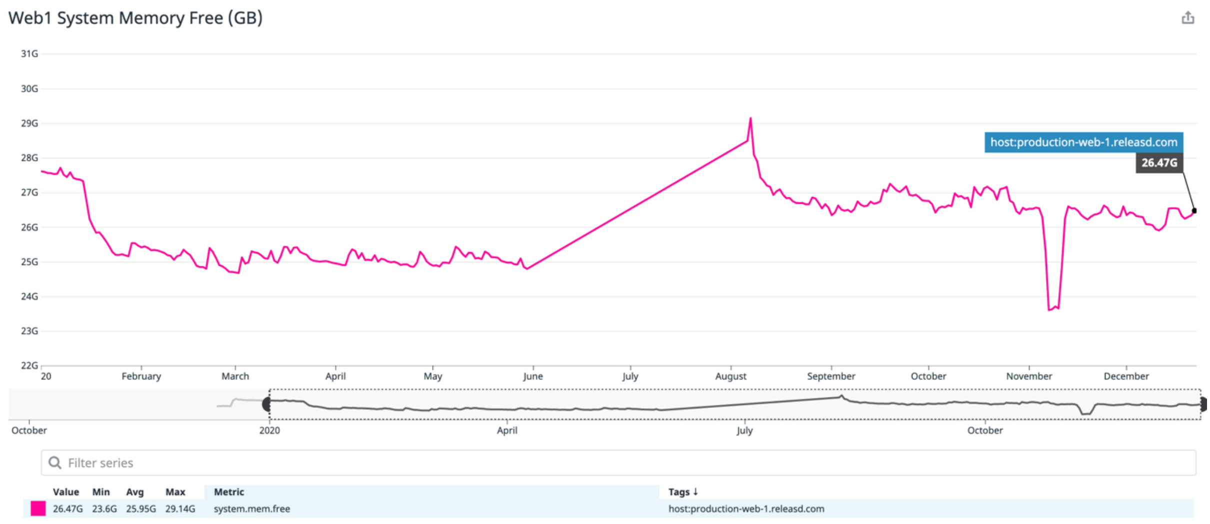
If we look at the historical data for CPU and RAM usage since Jan 2020 we can see that this is significantly over provisioned.



CPU %

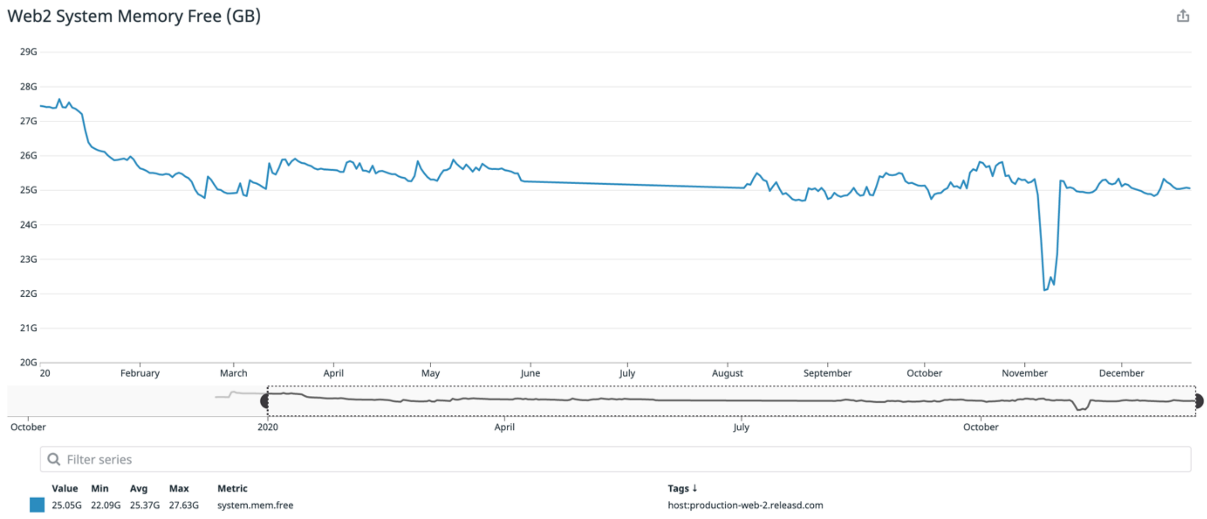
■ i-092b69073d09fbc1c (production-web-1 20191114) ■ i-0602e1f609a103e62 (production-web-2 20191114)

CPU Usage on Web-Server-1 & 2



Web1 System Memory Free (GB)

	Value	Min	Avg	Max	Metric	Tags ↓
■	26.47G	23.6G	25.95G	29.14G	system.mem.free	host:production-web-1.releasd.com



Web2 System Memory Free (GB)

	Value	Min	Avg	Max	Metric	Tags ↓
■	25.05G	22.09G	25.37G	27.63G	system.mem.free	host:production-web-2.releasd.com

The maximum RAM usage is around 8GB.

A recommended step here is to reduce the size of this instance from a m3.2xlarge to a m5.xlarge with the following specifications:

16GB Ram
4 vCPUs
EBS only

This would reduce the running cost of the existing Production Web Servers from:

£978.36 per month
£11740.32 per annum

To

£228.44 per month
£2741.28 per annum

· **Payment-db-cluster**

If this is currently not being utilised, then the db-cluster should be removed. This will offer a saving of:
£18.49 per month
£221.88 per annum

If we move forward with all of the above actions, then the total potential saving would be:

£13,772 per annum
£1147.6 per month

· **Released-PDF-Service**

This PDF server is also over provisioned. It currently runs:

64GB Ram

16 vCPUs

If this could be lowered to a m5.2xlarge which as the following specs:

32GB Ram

8 vCPUs

Then there could be a further cost saving of

£200 per month

£2400 per annum

This would give a total potential saving of

£1347.66 per month

£16172 per annum

· **Use reserved instance pricing.**

You are already aware of the benefits of reserved pricing. it is worth utilising reserved pricing in your business if you know that you will need the instance(s) for a set amount of time. This would potentially offer a further saving of up to 30%.

Considerations

I have noticed a few points within the architecture that I believe should be reviewed and possibly amended.

1. There are a large number of VPC/Subnets and Internet Gateways. If these are not all-in use, I suggest these be removed. This is not just spring cleaning; Internet Gateways have a cost.

2. Your two production instances are behind an Application Load Balancer that shares the traffic between the two instances. However, these instances are not part of an Auto Scaling Group. This means that if one of your instances was to go down it would not automatically replace itself. This would need to be done manually. Also, if you experience a surge in demand your instances will not automatically react and "scale-out" These are set to a specific size and shape and cannot change until manually changed.

   This is a naïve setup. Part of the advantage of cloud computing is the elasticity – or, put simpler – the ability to scale in and out to the current level of demand.

   I would strongly recommend you consider utilising auto-scaling groups to accommodate this. This would allow your webservers to deploy more when it is required and less when it is not needed.

   Currently you only have two webservers regardless of the level of traffic or health of these instances.

3. Before upgrading any instances it is worth considering any implications this may have on application dependencies. Are there currently software packages that require a specific type of hardware to run and is that why you are using older generation systems?

Suggested Next Actions

I would suggest that the following actions should take place:

1. If you just wanted to change your existing servers to the most recent generation equivalent this is a pretty quick fix. It essentially just requires a backup and then a stop and redeployment of the existing server. This means you could start saving costs almost immediately. This action alone could potentially save you over £500-700 per month with little to no extra configuration needed.

   It is recommended that you perform this action at an out of hours or seasonal time to avoid disruption to your existing customers. There will be some downtime as a result of restarting servers.

   If you utilise reserve pricing in combination with updating servers to the current generation you may see a reduction in EC2 cost towards £900 per month.

2. Understand if the staging environment is needed. If not, remove/stop until required. This is another quick resolution and cost saving measure. This should have little to no impact on the existing infrastructure.

3. Discuss with developers on any particular software requirements that may impede potential hardware changes should you wish to consider adjusting capacity.

4. Consider the proposed reductions of server sizes. Create any backups required before applying the new sizes.

5. Test the new size instance and ensure they are working as expected.

6. Implement an auto-scaling group that will change how many instances are deployed depending on the traffic. This way, if you do experience a significant increase in traffic you will not need to experience downtime.

7. Create a Disaster Recover plan to ensure business continuity. I cannot currently understand from your setup what your plan is and I am concerned that if your architecture was to go down it would take you potentially hours to recover.

8. The most recent instance snapshot for Ireland was taken in December 2019. This means that if you were to recover from a disaster then your latest backup would potentially be from 2019. This seems dangerous. Do you have some other form of recovery?

Disaster Recovery

Downtime could jeopardise your business, so having a clear understanding of how much an outage could cost you is imperative. Many businesses underestimate this cost.

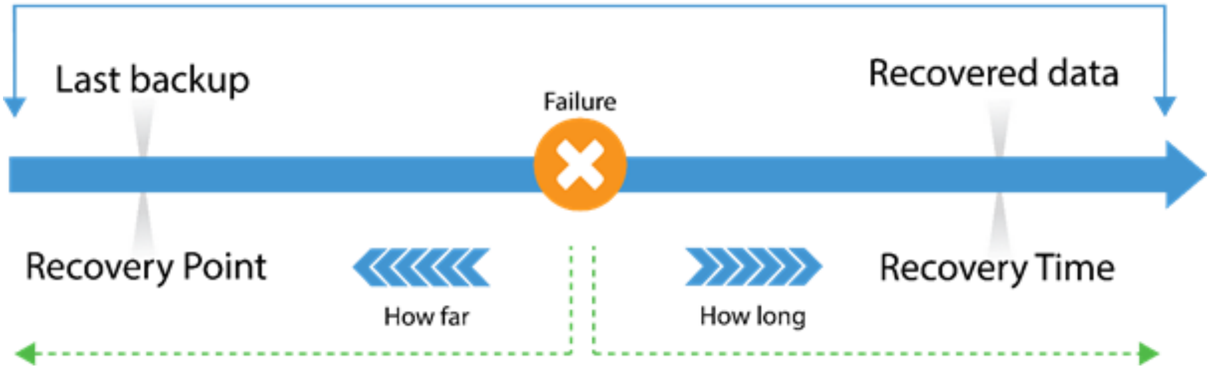
A disaster recovery plan is a formal document created by an organization that contains detailed instructions on how to respond to unplanned incidents such as natural disasters, power outages, cyber-attacks and any other disruptive events.

The plan contains strategies on minimizing the effects of a disaster, so an organization will continue to operate – or quickly resume key operations. Disruptions can lead to lost revenue, brand damage and dissatisfied customers. Therefore, a good disaster recovery plan should enable rapid recovery from disruptions, regardless of the source of the disruption.

Recovery Time Objective

The Recovery Time Objective (RTO) is the duration of time and a service level within which a business process must be restored after a disaster in order to avoid unacceptable consequences associated with a break in continuity.

In other words, the RTO is the answer to the question: "How much time did it take to recover after notification of business process disruption? "



Recovery Point Objective

The recovery point objective (RPO) is the age of files that must be recovered from backup storage for normal operations to resume if a computer, system, or network goes down as a result of a hardware, program, or communications failure.

The RPO is expressed backward in time (that is, into the past) from the instant at which the failure occurs, and can be specified in seconds, minutes, hours, or days. It an important consideration in disaster recovery planning (DRP).